RIVER DISCHARGE ANALYSIS OF

DANUBE AND ISAR

Abhijeet Parida - 03679676 and Nikhil Agarwal - 03681647

Technische Universität München

Contents

1. Introduction

For better prevention of flood events, thus planning protection measures and mitigation actions it is important to analyze the discharge of rivers and the regarding water height, e.g. in order to estimate how likely a critical water height could be exceeded. Daily measurements were given for the discharge of the two rivers Isar and Danube from 1925/11/01 to 2017/05/21 as they can be downloaded from Bayerisches Landesamt für Umwelt (www.gkd.bayern.de). The measurements are given for the gauge stations Plattling (Isar) and Hofkirchen (Danube). The estuarial area where the Isar River enters the Danube River lies in the commune Moos south of Deggendorf and east of Plattling.

2. Descriptive Statistical Analysis

Descriptive statistics include the numbers, tables, charts, and graphs used to describe, organize, summarize, and present raw data. Descriptive statistics are most often used to examine[1]:

- **Central tendency** (location) of data, i.e. where data tend to fall, as measured by the mean, median, and mode.
- **Dispersion** (variability) of data, i.e. how spread out data are, as measured by the variance and its square root, the standard deviation.

2.1 Central Tendency

Measures of Central Tendency indicate the middle and commonly occurring points in a data set. **Mean** is the average, the most common measure of central tendency. The mean may not always be the best measure of central tendency, especially if data are skewed. **Median** is the value in the middle of the data set when the measurements are arranged in order of the magnitude. **Mode** is the value occurring most often in the data.

The Table Tab 1-3. Shows a comparision of the central tendencies of both rivers.

RIVER	MEAN DISCHARGE (MEAN)	MEAN DISCHARGE (MAX)	MEAN DISCHARGE (MIN)
DANUBE	638.16	666.26	611.03
ISAR	172.3	185.87	159

Tab 1. Means of various discharges of in m^3/s Danube and Isar

RIVER	MODE DISCHARGE (MEAN)	MODE DISCHARGE (MAX)	MODE DISCHARGE (MIN)
DANUBE	524	1010	1010
ISAR	121	132	156

Tab 2. Modes of various discharges of in m^3/s Danube and Isar

RIVER	MEDIAN DISCHARGE (MEAN)	MEDIAN DISCHARGE (MAX)	MEDIAN DISCHARGE (MIN)
DANUBE	559	538	580
ISAR	156	143.5	159

Tab 3. Median of various discharges of in m^3/s Danube and Isar

2.2 Dispersion

Measures of Dispersion indicate how spread out the data are around the mean. **Variance** is expressed as the sum of the squares of the differences between each observation and the mean, which quantity is then divided by the sample size . **Standard deviation** is expressed as the positive square root of the variance. The standard deviation is used when expressing dispersion in the same units as the original measurements.

The Table Tab 4-5. Shows a comparision of the dispersions of both rivers.

RIVER	VARIANCE DISCHARGE (MEAN)	VARIANCE DISCHARGE (MAX)	VARIANCE DISCHARGE (MIN)
DANUBE	91411	1.0473e05	78842
ISAR	172.3	185.87	159

Tab 4. Variance of various discharges of Danube and Isar

RIVER	STD. DEV DISCHARGE (MEAN)	STD. DEV DISCHARGE (MAX)	STD. DEV DISCHARGE (MIN)
DANUBE	302.34	323.63	280.79
ISAR	69.045	77.072	62.606

Tab 5. Standard Deviation of various discharges of Danube and Isar

2.3 Range, Min and Max[2]

A few more important parameters for the statistical visualization of the data are Range, Minimum and Maximum. The min is simply the lowest observation, while the max is the highest observation. Finding the min and max helps us understand the total span of our data. Sometimes it is also useful to use the min and max to calculate the range of a dataset. The range is a numerical indication of the span of our data.

The Tab. 6-8 ennumerates these properties of the given data.

RIVER	MIN DISCHARGE (MEAN)	MIN DISCHARGE (MAX)	MIN DISCHARGE (MIN)
DANUBE	0	0	0
ISAR	0	0	0

Tab 6. Min of various discharges in m^3/s of Danube and Isar

RIVER	MAX DISCHARGE (MEAN)	MAX DISCHARGE (MAX)	MAX DISCHARGE (MIN)
DANUBE	3450	3510	3320
ISAR	1100	1180	1050

Tab 7. Max of various discharges in m^3/s of Danube and Isar

RIVER	RANGE DISCHARGE (MEAN)	RANGE DISCHARGE (MAX)	RANGE DISCHARGE (MIN)
DANUBE	3450	3510	3320
ISAR	1100	1180	1050

Tab 8. Range of various discharges of in m^3/s Danube and Isar

2.4 Quantiles[3], Quartiles[4] and Outliers[5,6]

The word quantile has no fewer than two distinct meanings in probability. Specific elements x in the range of a variate X are called quantiles, and denoted x .This particular meaning has close ties to the so-called quantile function, a function which assigns to each probability p attained by a certain probability density function $f = f(X)$ a value $Q_f(p)$ defined by

$$Q_f(p) = \{x : \Pr(X \le x) = p\}.$$

One of the four divisions of observations which have been grouped into four equal-sized sets based on their statistical rank. The quartile including the top statistically ranked members is called the first quartile and denoted Q_1. The other quartiles are similarly denoted Q_2, Q_3, and Q_4. For N data points with $N$of the form $4n + 5$ (for $n = 0$, 1, ...), the hinges are identical to the first and third quartiles.

The Tab. 9-11 enumerates these Quartiles of the given data of both rivers.

RIVER	QUARTILE 1 DISCHARGE (MEAN)	QUARTILE 1 DISCHARGE (MAX)	QUARTILE 1 DISCHARGE (MIN)
DANUBE	437	451	424
ISAR	126	135	117

Tab 9. Quartile 1 of various discharges in m^3/s of Danube and Isar

RIVER	QUARTILE 2 DISCHARGE (MEAN)	QUARTILE 2 DISCHARGE (MAX)	QUARTILE 2 DISCHARGE (MIN)
DANUBE	559	580	538
ISAR	156	169	143.5

Tab 10. Quartile 2 of various discharges in m^3/s of Danube and Isar

RIVER	QUARTILE 3 DISCHARGE (MEAN)	QUARTILE 3 DISCHARGE (MAX)	QUARTILE 3 DISCHARGE (MIN)
DANUBE	756	793	722
ISAR	200	215	185

Tab 11. Quartile 3 of various discharges in m^3/s of Danube and Isar

An outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution.

The outlier are just values too far from the Quartiles and number of outliers are tabulated in Tab 12.

RIVER	OUTLIERS IN DISCHARGE (MEAN)	OUTLIERS IN DISCHARGE (MAX)	OUTLIERS IN DISCHARGE (MIN)
DANUBE	565	554	527
ISAR	455	514	435

Tab 12. Outliers in the data of of various discharges in m^3/s of Danube and Isar

3 Probability Distribution

The probability distribution is a description of a random phenomenon in terms of the probabilities of events. Examples of random phenomena can include the results of an experiment or survey. A probability distribution is defined in terms of an underlying sample space, which is the set of all possible outcomes of the random phenomenon being observed. Every random variable has a probability distribution, which specifies the probability that its value falls in any given interval.

3.1 Identifying Distribution

In Matlab, pd = fitdist(x, distname) creates a probability distribution object by fitting the distribution specified by distname to the data in column vector x.

In Matlab, probplot(dist,y) creates a probability plot for the distribution specified by dist, using the sample data in y.

To visualize the fit of the specified distribution, the probability plot is examined for how closely the data points follow the fitted distribution line. If the specified theoretical distribution is a good fit, the points fall closely along the straight line.

3.2 Plotting PDF

The histogram of the data can be plotted in matlab using, hist(x, nbins) that creates a histogram bar chart of the elements in vector x. The elements in x are sorted into equally spaced bins along the x-axis between the minimum and maximum values of x. hist displays bins as rectangles, such that the height of each rectangle indicates the number of elements in the bin.

In Matlab, y = pdf(pd, x) returns the probability density function of the probability distribution object, pd, evaluated at the values in x.

The area under the fitted distribution is scaled with the area under the histogram.

3.3 Quantile-Quantile Plot (Probability Plot)

The quantile function is the inverse CDF, which returns the value x such that the cumulative probability at x is q.

A probability plot (Q-Q plot) is a graphical technique for comparing two data sets, either two sets of empirical observations, one empirical set against a theoretical set, or (more rarely) two theoretical sets against each other.

3.3.1 Difference in means, all else constant.

A pure difference in means looks like a constant offset from the identity line in the QQ plot. If the QQ plot is always below the identity line, that means that the variable plotted on X is larger than Y by a constant offset.

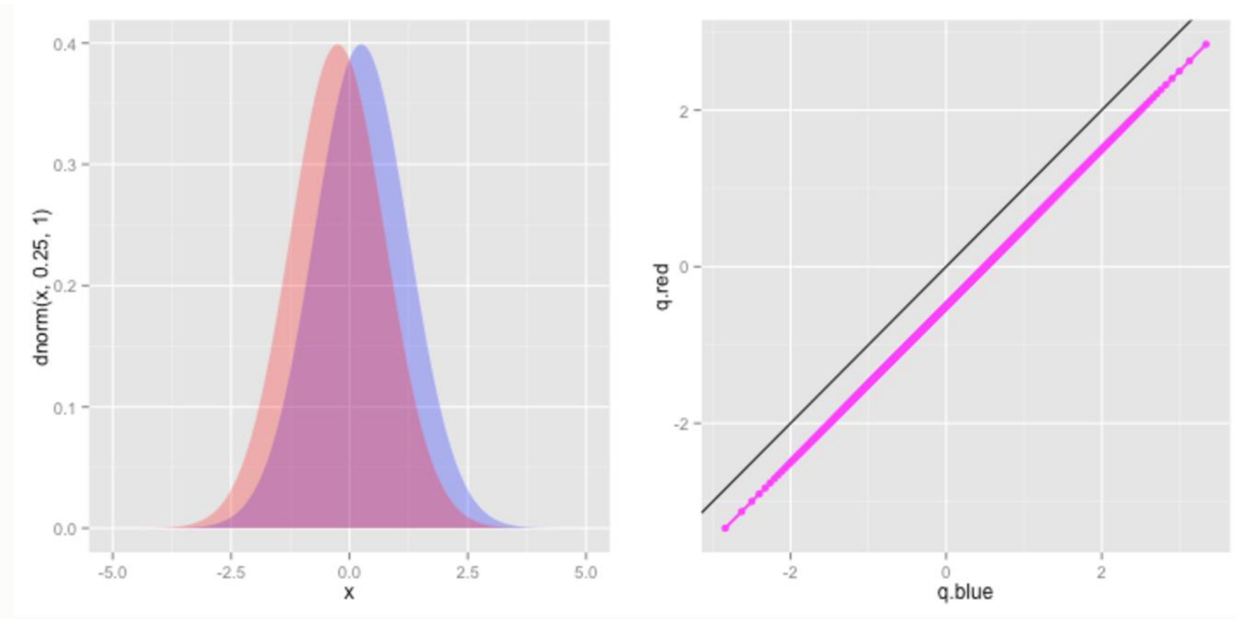


Fig.1: PDF and Q-Q Plot with 2 random variables with different means rest being the same

3.3.2 Difference in variance, all else constant.

A pure difference in variance will manifest as a slope that differs from 1. For instance, if X has the same mean, but higher variance than Y, then the upper quantiles (those close to 1) of X will be larger than those of Y (so the QQ plot will be below the identity line at the right of the plot), and the lower quantiles (those near 0) of X will be smaller than those of Y (so the QQ points will be above the identity line at the left of the plot). Consequently, if X has a larger variance than Y, then the QQ plot will have a slope less than 1; a slope greater than 1 indicates the opposite – that Y has a larger variance than X.

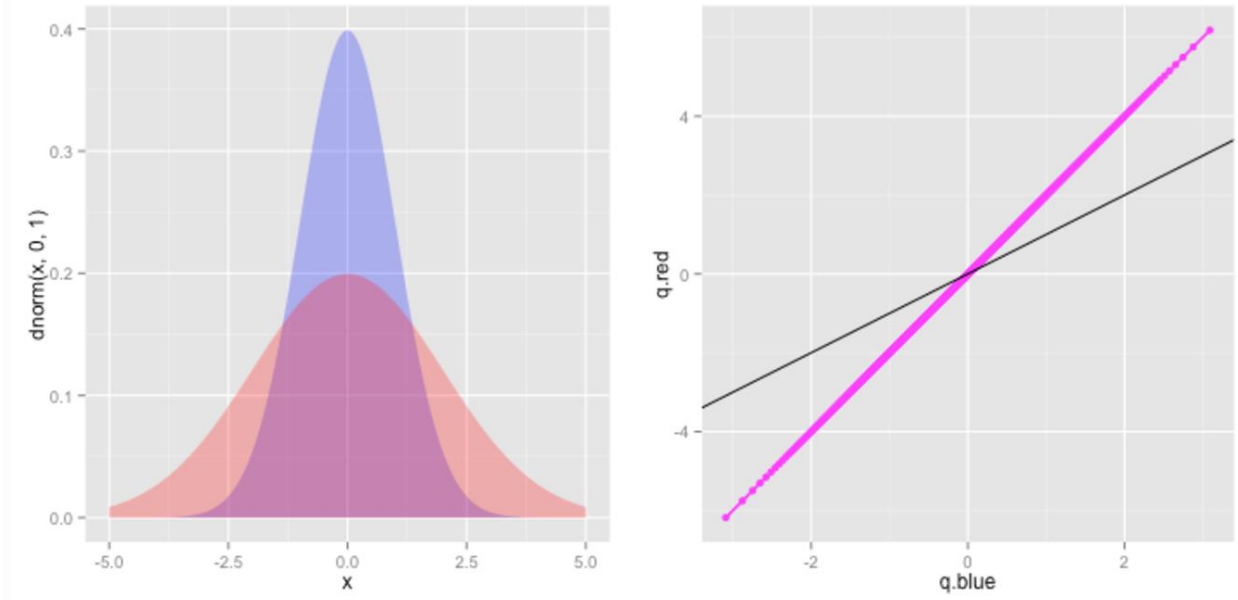


Fig.2: PDF and Q-Q Plot with 2 random variables with different variance and rest being the same

3.3.3 Difference in skew, all else constant.

A difference in skew, with constant mean and variance means that one distribution will have a heavier tail in one direction than the other. Thus, if X has a more positive skew than Y, then upper quantiles of X will be larger than those of Y, and the lower quantiles of Y will be smaller than those of X, but the middle quantiles of X will be roughly matched to those of Y. Consequently, we will get a curved shape: the qq plot will be below the identity line at the left and right of the plot, and at or above the identity line in the middle – a shape that is concave down. If X has a more negative skew, we will get a curved shape that is concave up.

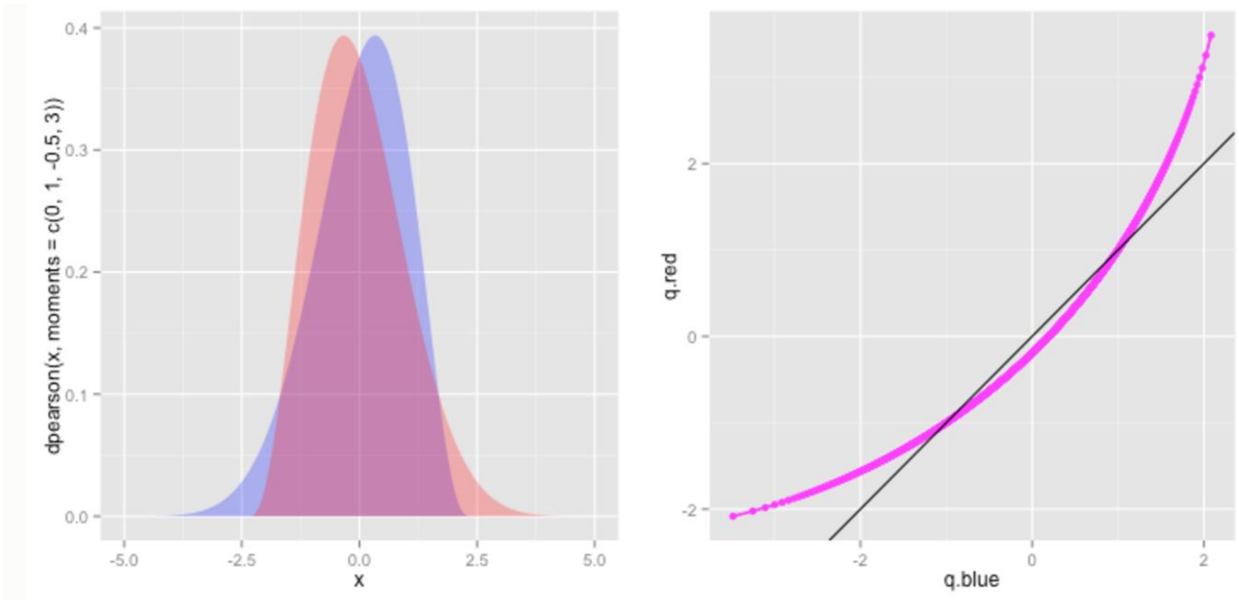


Fig.3: PDF and Q-Q Plot with 2 random variables with different skewness and rest being the same

3.3.4 Difference in kurtosis, all else constant

A difference in kurtosis with constant skew, variance, and mean would make the higher kurtosis distribution have heavier tails at the extremes, but lighter tails at less extreme points. Thus the QQ plot will look like a quadratic squiggle. If X has a greater kurtosis than Y, then its the extreme positive quantiles will be larger, and its extreme negative quantiles will be smaller than Y, thus the left and right ends of the qq plot will be above and below the identity line, respectively, but this pattern will reverse near the middle. (And the converse will be true if X has a lower kurtosis than Y.)

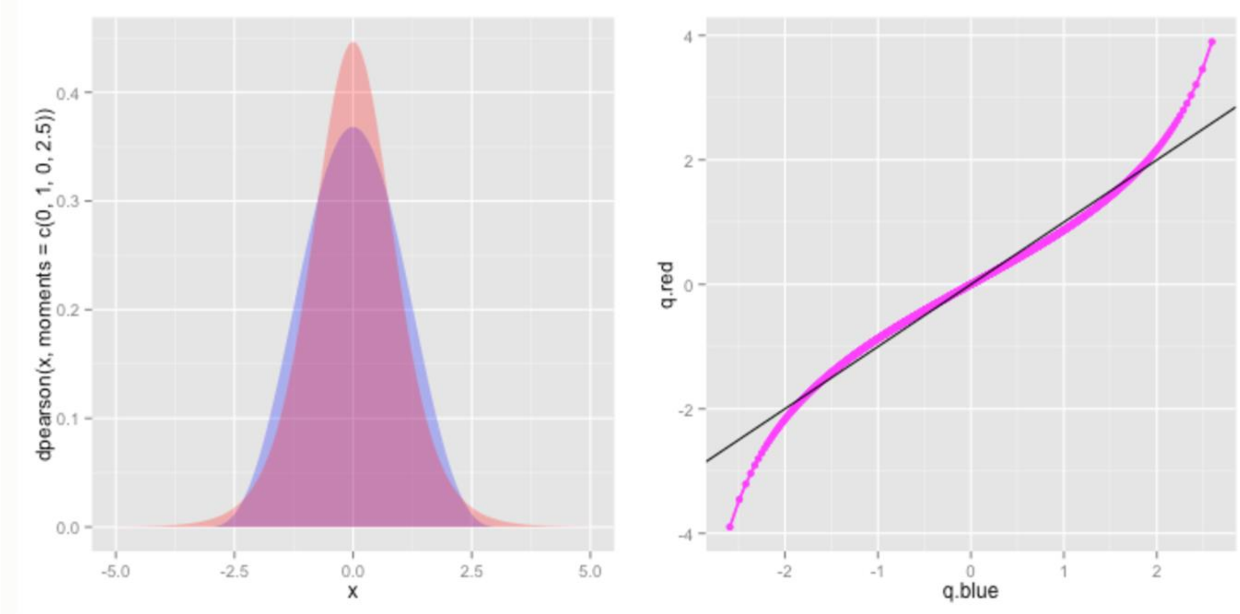


Fig.4: PDF and Q-Q Plot with 2 random variables with different kurtosis and rest being the same

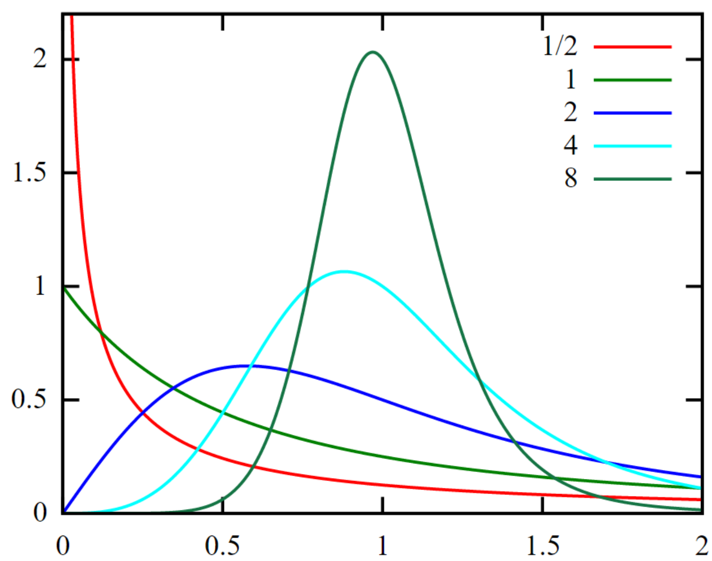
3.4 Quantitative Evaluation of Distribution fit

The distribution fit could be evaluated quantitatively by finding L2 norm of the errors of the quantile – quantile plot or the PDF plot.

3.5 Log-logistic Distribution

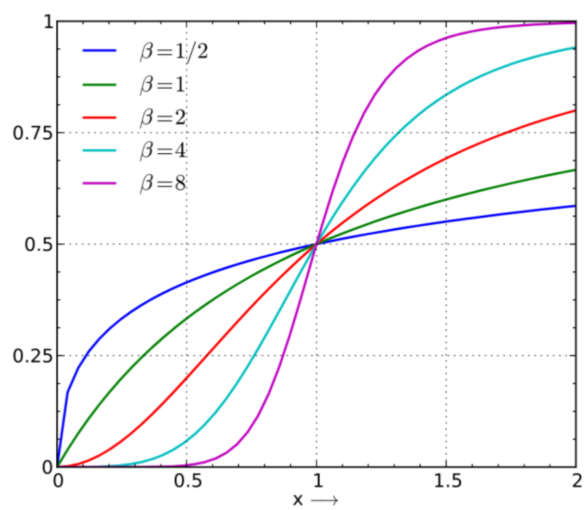
Log-logistic distribution (known as the Fisk distribution in economics) is a continuous probability distribution for a non-negative random variable. It is used in survival analysis as a parametric model for events whose rate increases initially and decreases later, for example mortality rate from cancer following diagnosis or treatment. It has also been used in hydrology to model stream flow and precipitation, in economics as a simple model of the distribution of wealth or income, and in networking to model the transmission times of data considering both the network and the software.

The log-logistic distribution is the probability distribution of a random variable whose logarithm has a logistic distribution. It is similar in shape to the log-normal distribution but has heavier tails. Unlike the log-normal, its cumulative distribution function can be written in closed form.



α=1, values of β as shown in legend

Fig.5: Probability density functions: Log-logistic Distribution



α=1, values of β as shown in legend

Fig.6: Cumulative distribution function: Log-logisctic Distribution

The parameter α>0 is a scale parameter and is also the median of the distribution. The parameter β>0 is a shape parameter. The distribution is unimodal when β>1 and its dispersion decreases as β increases.

The probability density function is

$$f(x;\alpha,\beta) = \frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{(1+(x/\alpha)^{\beta})^2} \tag{1}$$

The quantile function (inverse cumulative distribution function) is :

$$F^{-1}(p;\alpha,\beta) = \alpha\left(\frac{p}{1-p}\right)^{\frac{1}{\beta}} \tag{2}$$

4. Monte Carlo Simulations

Monte Carlo methods (or Monte Carlo experiments) are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. Their essential idea is using randomness to solve problems that might be deterministic in principle. They are often used in physical and mathematical problems and are most useful when it is difficult or impossible to use other approaches.[7]

4.1. How Monte Carlo Simulation Work?[8]

Monte Carlo simulation performs risk analysis by building models of possible results by substituting a range of values—a probability distribution—for any factor that has inherent uncertainty. It then calculates results over and over, each time using a different set of random values from the probability functions. Depending upon the number of uncertainties and the ranges specified for them, a Monte Carlo simulation could involve thousands or tens of thousands of recalculations before it is complete. Monte Carlo simulation produces distributions of possible outcome values.

4.2. Monte Carlo Simulation to predict flooding of Isar

The following steps need to carried out:

- First the distribution that fits the maximum daily discharge of the Isar has to be obtained.
- Then random values of this distribution with required parameters are generated for a initial one thousand for the seeds for the discharge of the Isar.
- The formula that correlates the Discharge for the river and the height of the river is obtained from the rating curve[9] as:

$$Q = C_r(G - a)^{\beta} \tag{3}$$

Where, Q= Discharge(m^3/s)

Cr=Rating Constant

a= height at zero discharge

G=height of the river

ß=gauging parameter (range of 2-3 under section control, and 1-2 under channel control.)

- The value of ß was chosen to be 2 as we are not aware of the section control or channel control. The value of a was chosen to be the ETA value in the resource[10]. The unknown Cr was determined by using the value from the Ablusstafel[10].
- From the seeds the various height from the formula was obtained
- Then the probability of flood was detected by determining the value by using the critical value of the h_{cr}=4.7m.
- The probability was used to generate the minimum of simulations to stay within the Coefficient of Variation within 10%
- The Monte Carlo steps was recreated for the new sample size.

5. Results

From the time series plot seen in Fig.1. A mere visualization can help up that there may be linear dependencies in the data sets from both gauge station by matching crest with crest and trough with trough. This shift is most likely due to the Danube being much larger river.

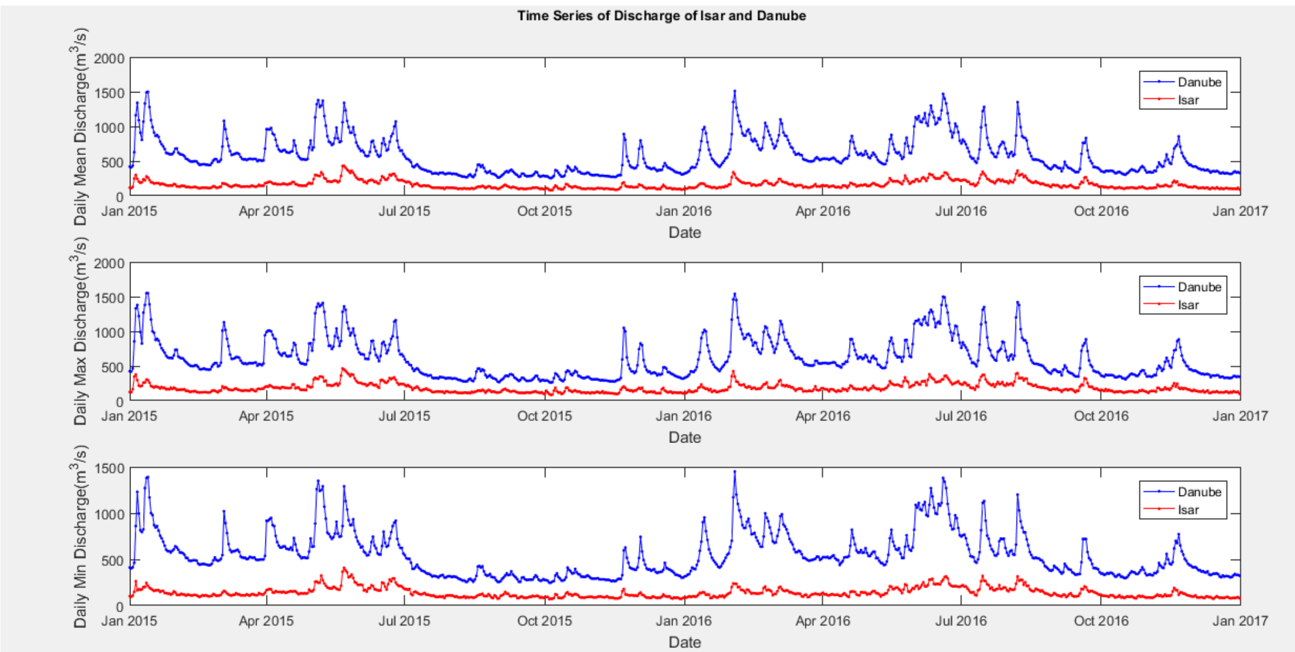


Fig. 1 shows the time series of various discharges from Jan 2015-Jan2017

The Fig. 2 and Fig 3 the histogram of the discharges of the rivers of Isar and Danube. From visualization it is possible that with the same bin size we can say the data follows from the same distribution. Also the outliers can be seen are mostly the outlier which are too ;arge from the Q3 of each distribution.

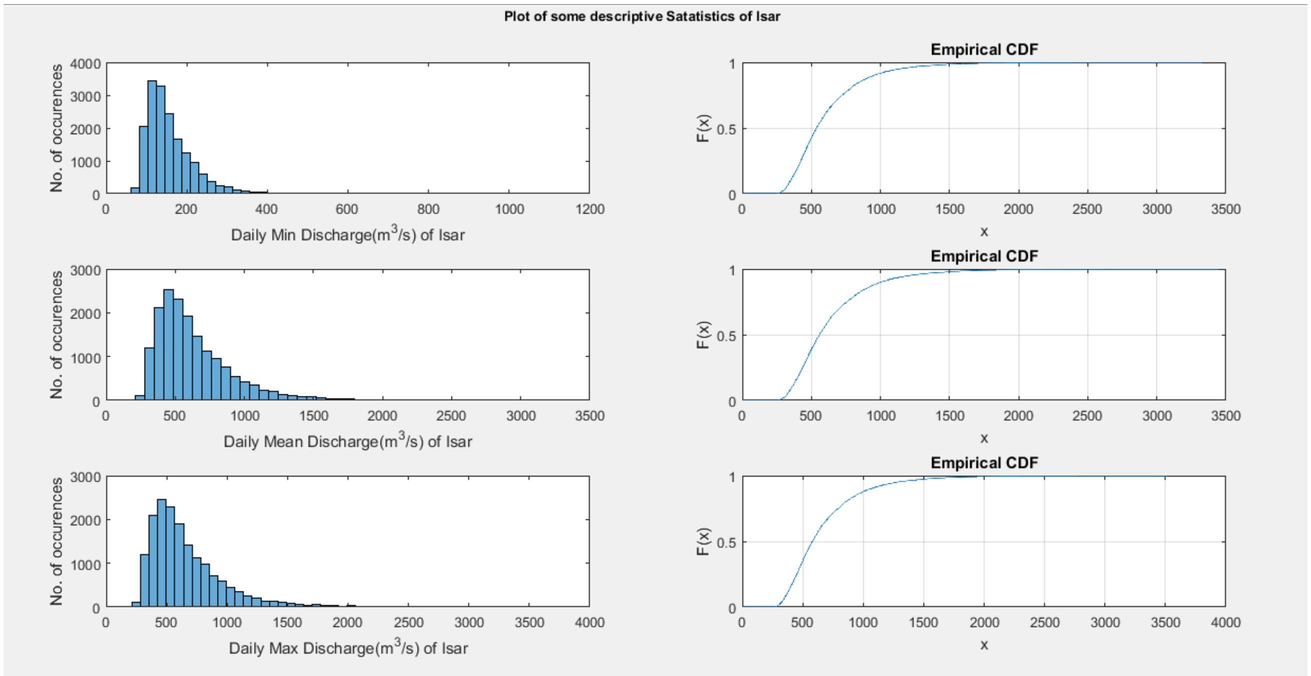


Fig. 2 Histogram of Discharges of Isar and the corresponding CDF

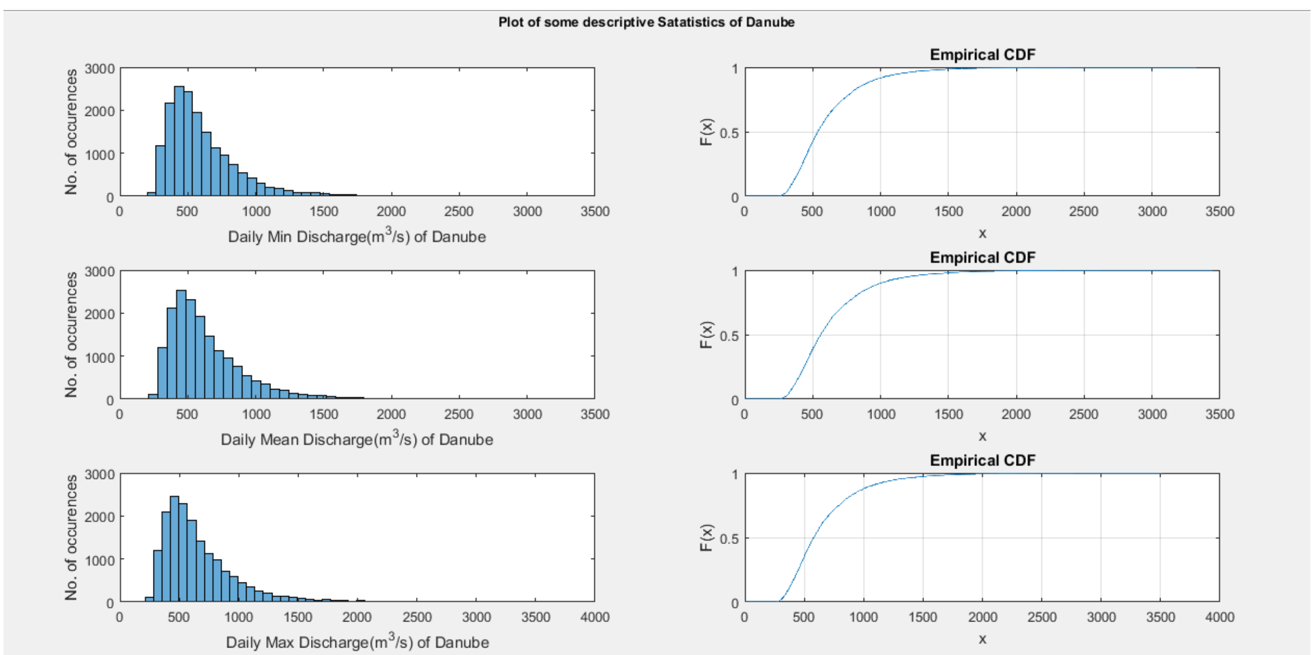


Fig. 3 Histogram of Discharges of Danube and the corresponding CDF

From the data analysis, it is found that the annually the maximum river height at Platting occurs most likely in the months of June to August based on the analyses during the period 1971 to 2016.

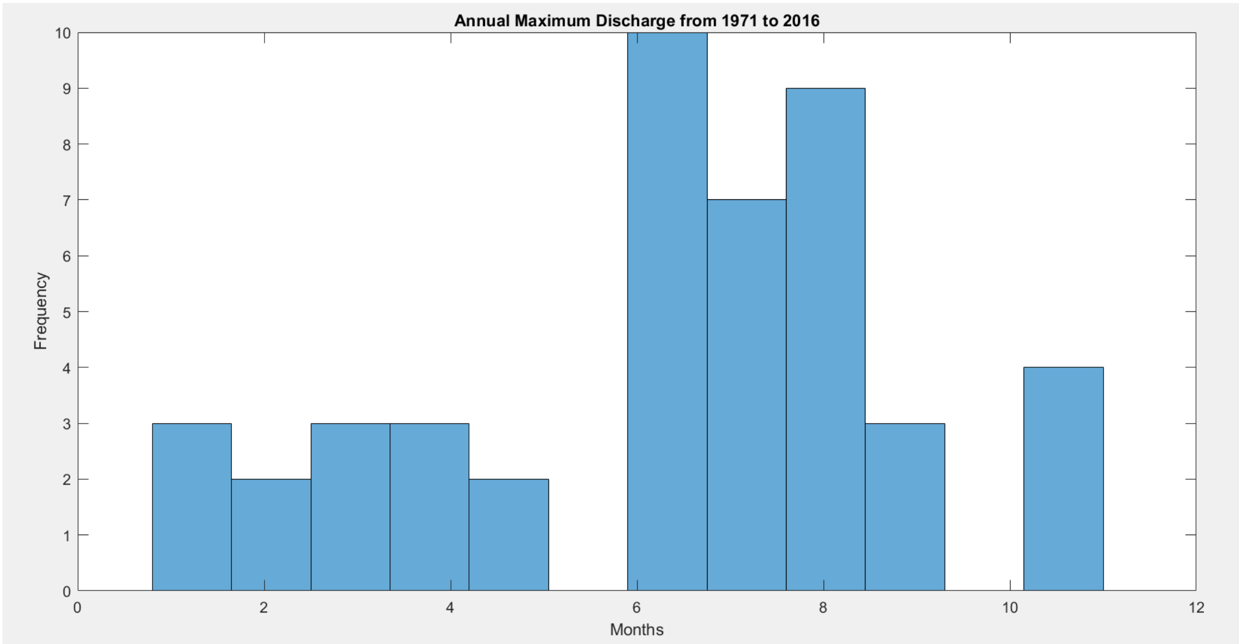


Fig.4: Annual Maximum Discharge from 1971 to 2016 at Plattling

Based on the comparison of Probability distribution plot superimposed on the histogram and also from the Quantile-Quantile plot we found that, log logistic distribution is better representative of the data for Annual Mean and Annual Maximum Discharge of Isar measured at Plattling.

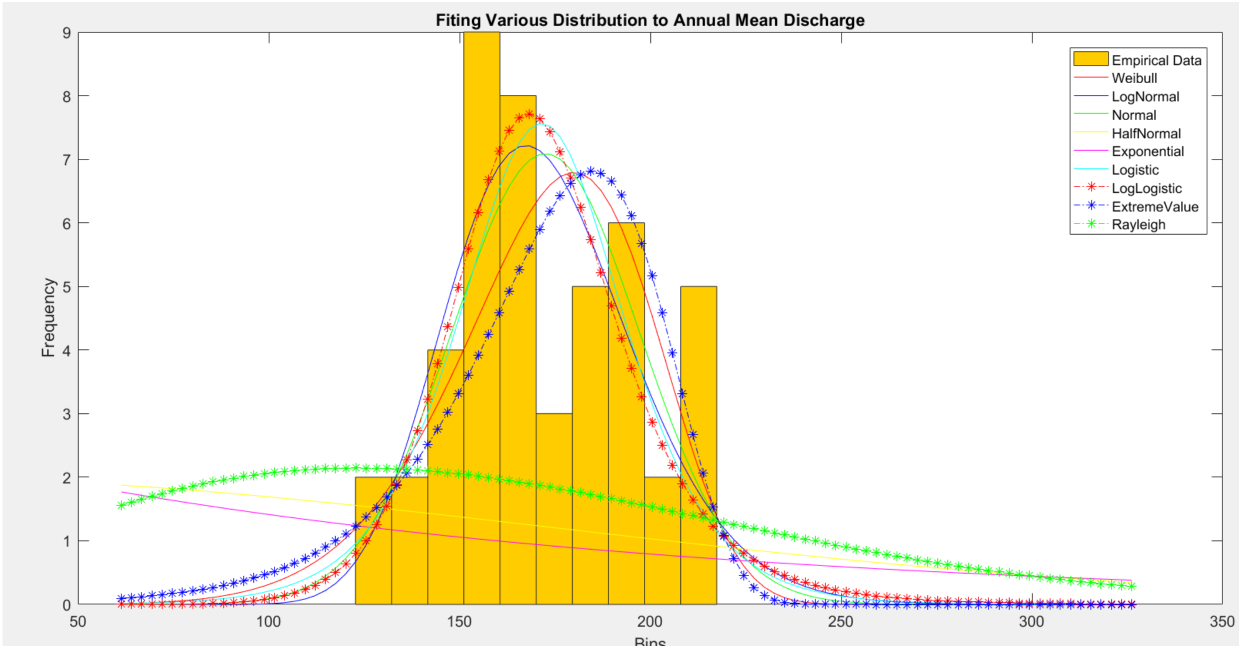


Fig.5: Fitting Various Distributions to Annual Mean Discharge at Plattling

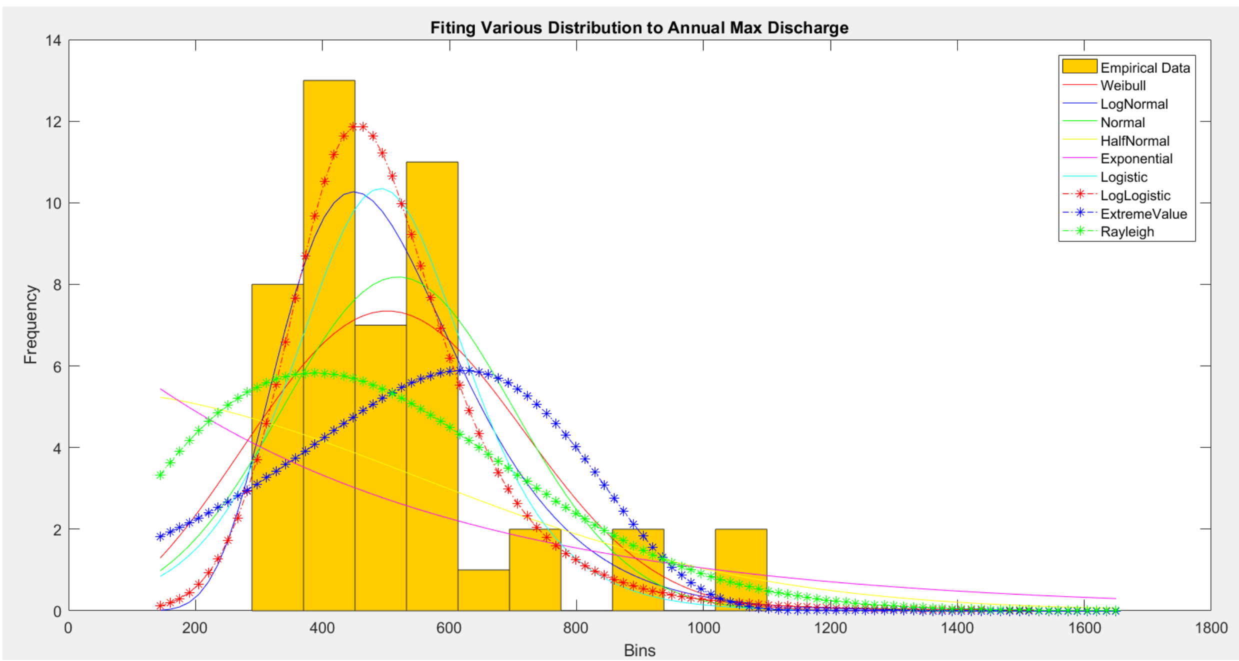


Fig.6: Fitting Various Distributions to Annual Max Discharge

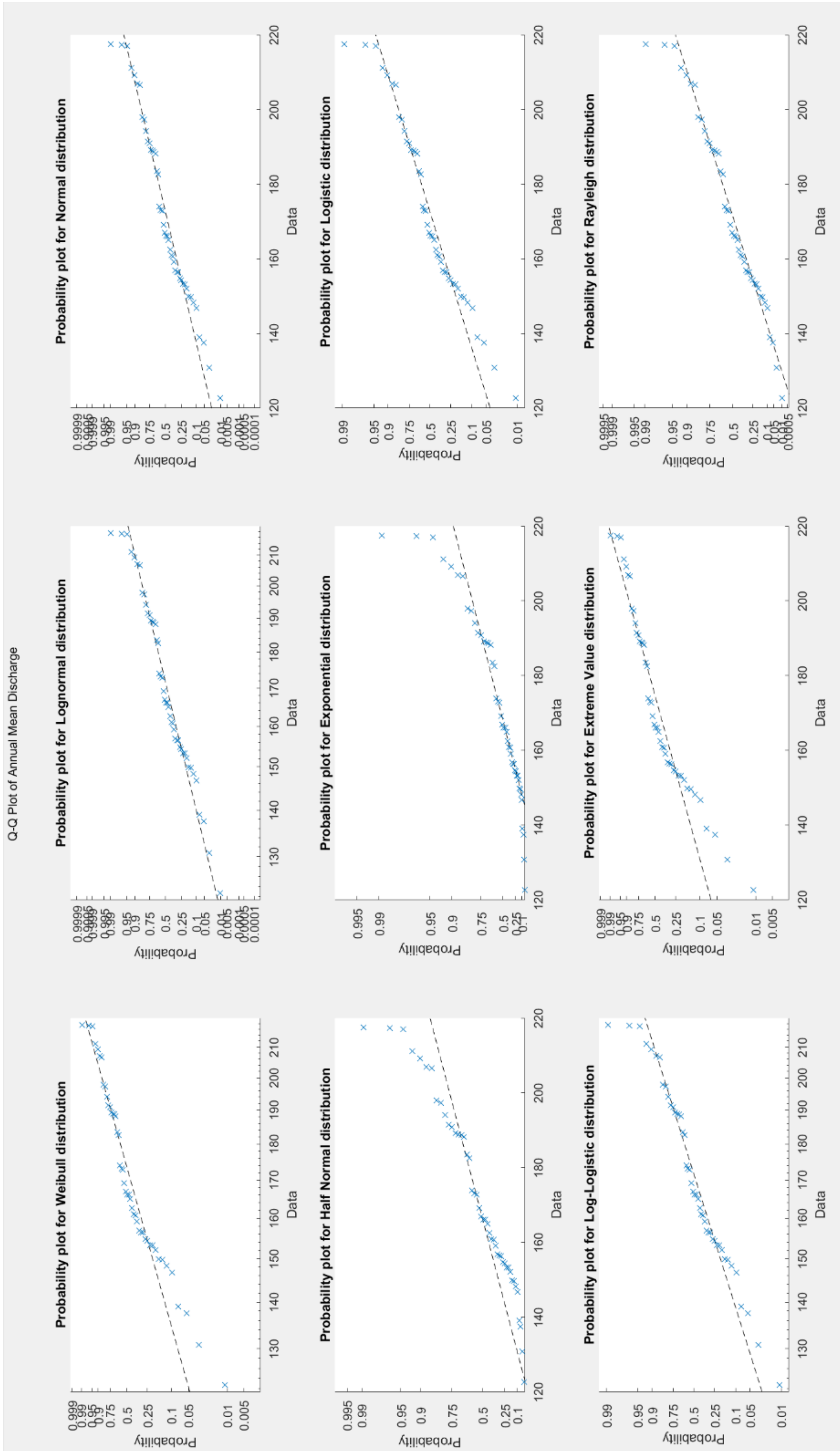


Fig.7: Q-Q Plot of Annual Mean Discharge at Plattling from 1971 to 2016

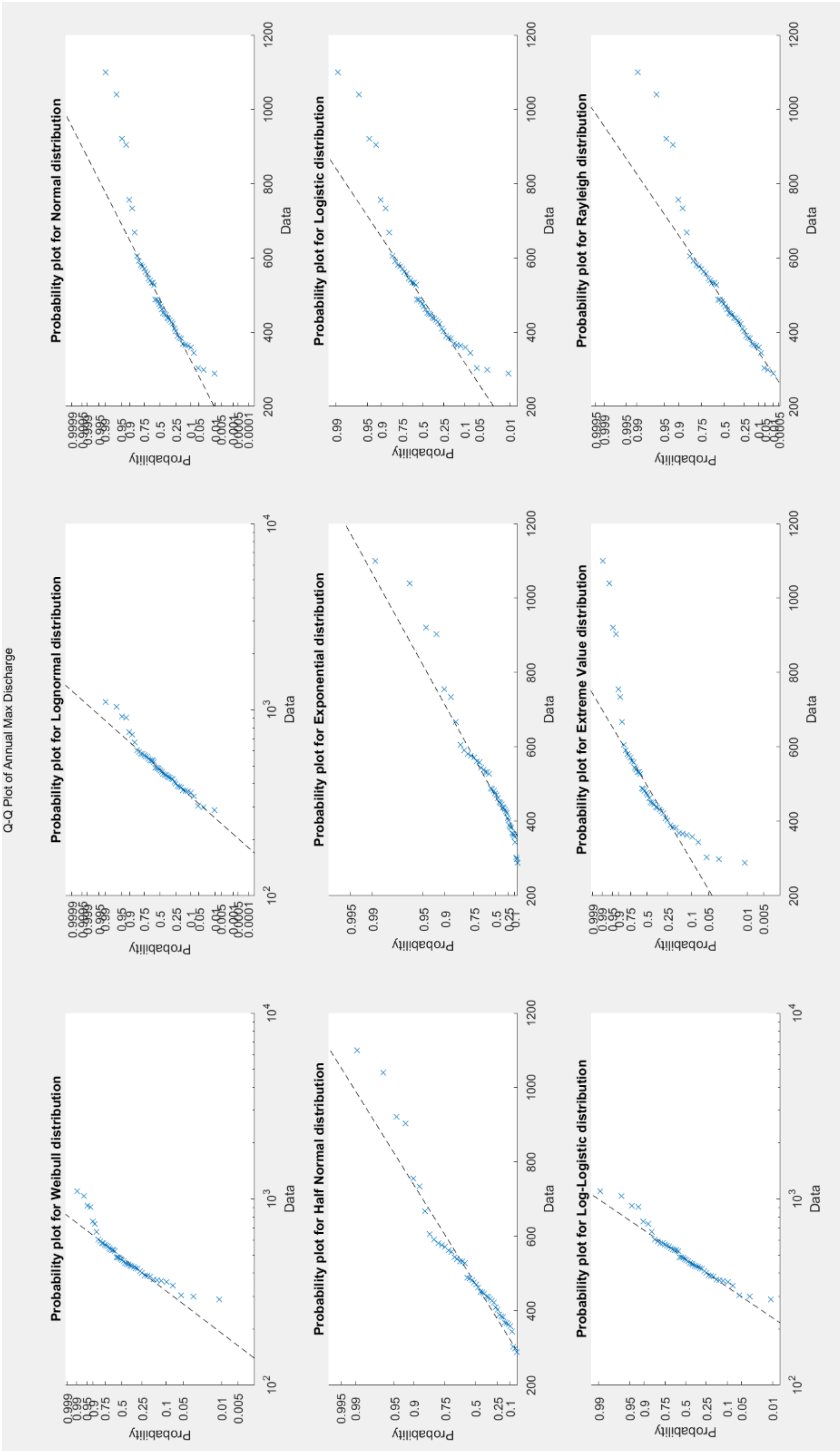


Fig.8: Q-Q Plot of Annual Maximum Discharge at Plattling from 1971 to 2016

The Parameters of Log-Logistic Distribution for Annual Mean Discharge are given in Tab. 13.

Log mean	Log scale parameter
5.13899	0.0833568

Tab. 13 The Parameters of Log-Logistic Distribution for Annual Mean Discharge at Plattling

The Parameters of Log-Logistic Distribution for Annual Maximum Discharge are given in Tab. 14.

Log mean	Log scale parameter
6.17822	0.167287

Tab. 14 The Parameters of Log-Logistic Distribution for Annual Maximum Discharge at Plattling

The result for one instance of the Monte Carlo Simultion to obtain the height are given in Tab. 15.

Initial Seed Size	Initial Probability of Flooding	Final Seed Size	Final Probability of Flooding
1000	0.002	50001	0.00004±10%

Tab. 15 an Instance of the Montecarlo Simulation

The Fig. 13 shows the variation of the height of the river obtained thru Monte Carlo Simulation and its variation of the parameters of the formula. It shows that the results are very sensitive to these parameters. This results in small changes in parameter giving large changes in the height of the river.

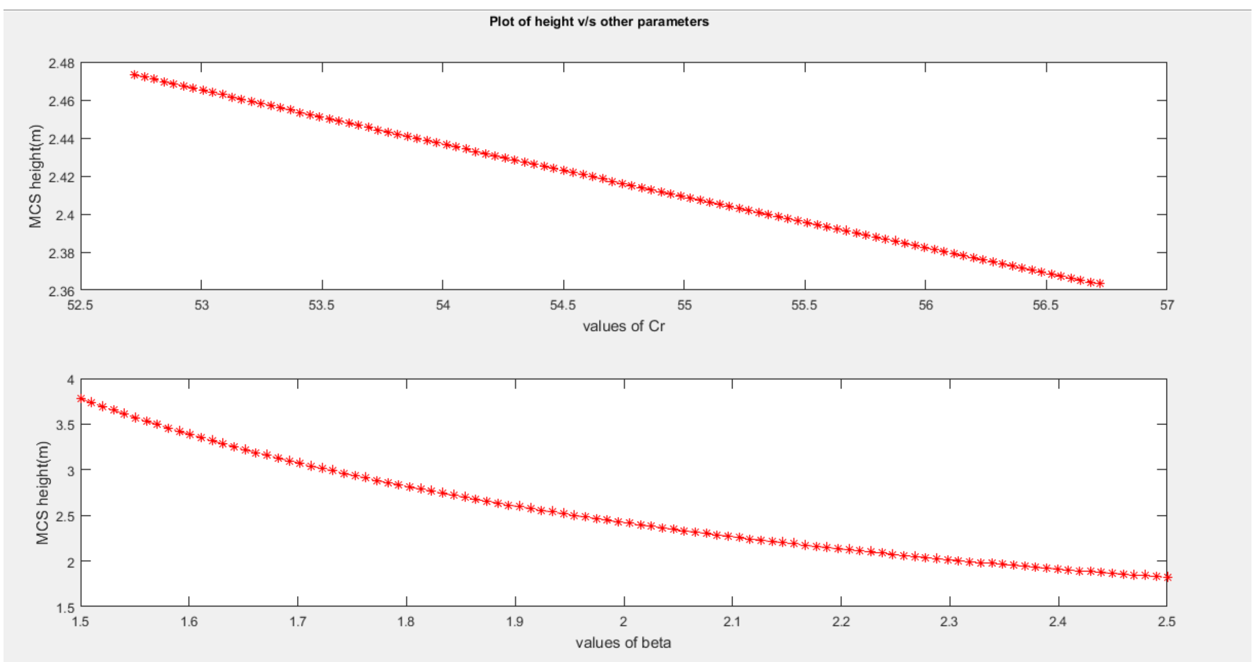


Fig. 9 Variation of height vs beta and Cr

4. Conclusion

The chances of occurrence of maximum discharge at Plattling is in the months from June to August. Probability Distribution of Annual Mean and Maxmum Discharge of Isar measured at Plattling is Log-logistic distribution.

The probability of flooding of Isar at Platting is 0.00004±10%. The results are very sensitive to the curve ratings.

5. References

1. http://vulstats.ucsd.edu/notes/visualization/qq-plots.html
2. http://www.nedarc.org/statisticalHelp/basicStatistics/minAndMax.html
3. http://mathworld.wolfram.com/Quantile.html
4. http://mathworld.wolfram.com/Quartile.html
5. https://en.wikipedia.org/wiki/Outlier
6. http://www.wikihow.com/Calculate-Outliers
7. https://en.wikipedia.org/wiki/Monte_Carlo_method
8. http://www.palisade.com/risk/monte_carlo_simulation.asp
9. https://en.wikipedia.org/wiki/Rating_curve
10. http://www.hnd.bayern.de/pegel/isar/plattling-16008506/abflusstafel